informatica

White Paper

Data Quality Management:

Beyond the Basics

While every attempt has been made to ensure that the information in this document is accurate and complete, some typographical errors or technical inaccuracies may exist. Informatica does not accept responsibility for any kind of loss resulting from the use of information contained in this document. The information contained in this document is subject to change without notice.

The incorporation of the product attributes discussed in these materials into any release or upgrade of any Informatica software product—as well as the timing of any such release or upgrade—is at the sole discretion of Informatica.

This edition published October 2014

Table of Contents

Introduction 2

Data is Not a Problem; It is An Opportunity 2

Research Reports the Data Explosion is Here to Stay 3

Why is Data Quality All the Buzz? 4

Faulty Data Results in Faulty Decisions 4

How Do Data Quality Problems Come About? 5

Data Quality Tools 5

Profiling Data Improves Efficiency 6

Standardizing Data Ensures Consistency. 7

Matching Multiple Data Records Reduces Duplication 8

Validating Third-Party Data Improves Accuracy and Enriches Data 8

Exploring Other Types of Data to Help Drive Results. 9

Improving Data Quality and Measuring Success 9

Conclusion 10

Introduction

Data is exploding. Big data. Real-time data. Consumer data. User-generated Internet data. Social data. Mobile data. Cloud-computing data. No matter how it is framed or defined, data is the lifeblood of all business. Data Quality Management (DQM) continues evolving as innovative solutions and new technologies enter the market place. The topic is hot with many questions to consider:

Is the data real-time? Is it accurate? How is it valuable to your business? Is it the right data? Is it actionable? Is it easy to integrate into both your internal and external business processes? Does it deliver value to your prospects, buyers, customer-facing employees and shareholders? Does it offer leadership and cross-functional teams new business intelligence, helping them make better informed, market-research based decisions? Does it provide quantifiable insights that executives can use to create strategic, forward-thinking decisions for building brand awareness and ROI?

Not all data is equal. While all data may be important, it is critical to evaluate and prioritize data based on quality, usability and total value to your business mission, goals and objectives.

The goal of this white paper is to:

- Educate you about the rapid growth of data quality
- Explain how managing data quality improves your business
- Challenge your data collection and integration methods
- Share knowledge and B2B best practices around DQM
- Provide insights into different types of approaches and tools to generate a strong ROI.

Data is Not a Problem; It is An Opportunity

Data is information. As such, data is mission critical. Basing decisions on accurate, high-quality data contributes to solid business decisions and offers valuable information, which helps you better understand your customers and target audiences. Data is a valuable asset and a capital investment. When managed correctly, data quality minimizes risk and maximizes ROI.

Every business faces a common problem. In today’s global economy, new information is constantly emerging as data moves from one person, to many people, to thousands of peer influencers, to potentially millions of new and existing customers. In some cases, new information even finds its way into communities and nations that businesses never thought possible.

Research Reports the Data Explosion is Here to Stay

Data quality, DQM and data communications are not abstract concepts.

However, the sheer amount of data that exists is overwhelming.

Consider these statistics:

- Businesses spent $130 billion on Short Messaging Services (SMS) in 2009, $875 million on data quality alone in 2012. It is estimated that by 2020 this spend will increase to $241 billion on data quality infrastructure, platform and software. (SOURCES: Mashable, IDC, Gartner, EMC, ABI Research).
- For enterprises, "Big data," typically refers to unstructured, variable data. The term means Petabytes (PB or one quadrillion bytes) and Exabyte (EB or one quintillion bytes) of data. There is enough data in one EB to fill more than 20 million, four-drawer filing cabinets with text. Over 2.5 quintillion bytes of new data appear on a daily basis. IBM reports 90% of the existing data in the world today is from the last two years alone. (SOURCE: blog.viralheat.com/2012/10/18/big-data-and-big-analytics)
    - Online data is exploding every day:
    - Email users send more than 144.8 billion emails (SOURCE: Radicati Group)
    - YouTube users upload 103,680 hours of new video (SOURCE: YouTube)
    - Facebook users share 2.5 billion content items (SOURCE: GIGAOM)
    - Consumers will spend an average of $619 million per day shopping online / eCommerce (SOURCE: Forrester Research)
    - Apple receives more than 46 million application downloads (SOURCE: Venturebeat)
    - Instagram users share 5.2 million photos per day (SOURCE: Instagram)
    - Foursquare users perform over 2.5 billion check-ins (SOURCE: Foursquare)
    - WordPress users publish more than 1 million posts (SOURCE: WordPress)
- By 2020, B2B and B2C online interactions will reach 450 million per day from both enterprise "managed" and consumer-generated data. (SOURCE: IDC) That is the equivalent of 880 billion GB, or almost 900 EB of data. The same scale would extend from Earth to 66,243 miles past the moon. (SOURCE: Wikibon)

Why is Data Quality All the Buzz?

Almost all software – whether utilizing infrastructure such as an internal server, the Web, an external data center, mobile phones, or digital devices – relies on data processing in one form or another.

In other words, the data and information it represents fuels the logic behind the application, similarly, to how food provides energy to the human brain. The logic - defined as the science that deals with the principles, reasoning and criteria of validity – provides the intelligence behind the application. Like a GPS navigation system, the logic paves the road and routes the driver to and from the desired destination. To work as intended, all the pieces or parts need access and the technical capability to receive, process, report accurate data, and interact between two or more systems.

If the data is not accurate, it is not valuable. Inaccuracies cause serious harm such as wasted time and resources, lost opportunities, poor customer service, and brand damage. They can even result in misguided, mission critical business decisions.

To prevent and avoid problems, data must be accurate, complete, current and arrive in a desirable, easy-to-use format.

Business Intelligence applications, for example, analyze data to help executives make decisions about the forward direction of a company. The mission-critical data stems from many factors, including relevant trends, consumer-purchasing patterns and other information organizations require.

Faulty Data Results in Faulty Decisions

The majority of businesses including retail, healthcare, pharmaceutical, telecommunications, and higher education industries, collect customer data. Data collection may occur at Point-of-Sale (POS), website forms, mobile applications, etc. The reasons why businesses collect data may vary, but ultimately all organizations desire to provide better, more personalized customer service to increase engagement and revenue.

If contact information such as name, address, email, or phone number is wrong, the intended communication will not reach the recipient. The customer relationship withers away.

For example, sending messages to inaccurate email addresses can result in serious consequences like being filtered to SPAM/junk folders, or worse being blacklisted or fined a penalty of up to $16,000 (per email) by the Federal Trade Commission (FTC).

Without proper customer data and the ability to reach consumers, companies cannot make sound business decisions. For instance, companies rely on data to form conclusions about entering a new geographic area based on customer buying patterns within a certain radius of a specific zip code. They invest millions of dollars to refurbish an existing or construct a new brick-and-mortar store. They later learn the data was inaccurate and face legitimate concerns about finances, or consequences with Wall Street analysts and investors.

How Do Data Quality Problems Come About?

Historically, when new system applications are developed, there is little if any focus on data quality. Typically, data quality-related issues begin to surface after use of the application and the lack of quality data starts negatively affecting the business (e.g. inaccurate or incomplete operational data, increased customer complaints).

DQM experts agree it is rare for businesses to consider and put quality controls in place early in the architecture lifecycle. Instead, data quality becomes an issue when prospective customers indicate they are using a certain system or platform configurations, causing multiple data quality-related problems.

Sometimes the issues are simple. For example, if data validation mechanisms are not in place when the information is originally collected, results are often inaccurate. Integrating notifications to validate a physical mailing address, telephone or email contact is a best practice.

Scenarios that are more complex occur when different people have built multiple systems at different times, for different purposes, each of which generates data in different formats. Cloud-based technology allows companies to combine, validate and verify the data in real-time so it becomes reliable, usable and actionable, helping businesses to achieve their goals.

Long-term solutions may even involve non-technology causes. For example, businesses can put proper incentives in place to ensure call center representatives collect the correct customer information the first time.

These real life examples often result in costly data quality issues, which significantly, and negatively, affect your organization, its processes, customers, employees and bottom line.

Fortunately, there are different approaches and data tools to consider as part of a comprehensive strategy for improving overall data quality assets.

Data Quality Tools

The market for data quality tools has increased in recent years as more organizations understand the impact of poor-quality data and seek solutions for improvement.

In addition to the overlap, integrating contact information and purchases across both mediums means understanding when customers prefer each channel, and as well as how to get the most out of your customer base across storefronts.

Emerging technologies have expanded the ways in which businesses use data to introduce new initiatives and increase revenue, elevate customer service expectations and complicate IT infrastructure, hosting solutions and data integration.

Product data – often driven by Master Data Management (MDM) initiatives – and financial data – largely driven by compliance pressures – are two areas driving demand for data quality tools.

Enterprises and SMBs also rely on data quality tools to:

1. Profile and analyze data to inform decision-making
2. Standardize data for consistency
3. Improve accuracy
4. Enrich and append data
5. Match, cross-reference, merge and integrate multiple/duplicate customer records
6. Explore and validate numbers or non-textual data

1. Profiling Data Improves Efficiency

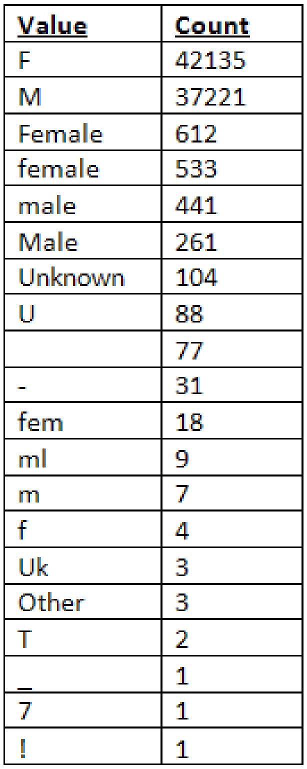
Before starting with data quality initiatives, businesses need to identify their underlying issues, evaluate the effects on different parts of the organization, and prioritize the problems discovered.

Areas to review include:

- What data is missing? Are certain fields within the records blank or null?
- What percent of duplicate customer data records exist? Can a matching criteria capability merge the existing data and prevent this from happening in the future?
- What types of data inconsistencies exist make the information difficult to analyze? Are you capturing the correct, most relevant data?

The answers to these questions will determine the most effective solutions to implement including integrating data quality metrics, developing a better understanding about how to use the data, and creating a data quality strategy to help minimize inaccurate data and maximize efficient use of data.

Consider this example. A business wants to capture a gender field within its master customer database. The data it collects produces the following report:



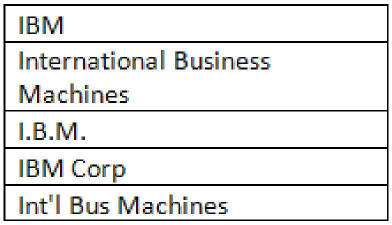
Value	Count
F	42135
M	37221
Female	612
female	533
male	441
Male	261
Unknown	104
U	88
	77
-	31
fem	18
ml	9
m	7
f	4
Uk	3
Other	3
T	2
_	1
7	1
!	1

What does this mean? Either this data did not have proper filters in place up front, or it came from multiple places. The result is inconsistency. If this data provides information to address letters with salutations (e.g. Mr. or Ms.), the algorithms need to account for these variations within the logic, or it will generate inaccurate results. This example also produced an "unknown," which means it is an empty value and therefore not useful.

2. Standardizing Data Ensures Consistency

One of the most common data quality problems businesses face is data consistency, as underscored in the previous gender example. However, the consistency issue affects all types of data, causing a myriad of issues.

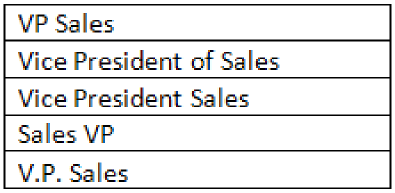
Another common example comes from businesses with call centers. They require representatives to collect a customer's "place of employment." The results are inconsistent:



IBM
International Business Machines
I.B.M.
IBM Corp
Int'l Bus Machines

The data results above present a considerable challenge in determining how to target customers from IBM, how many customers work at IBM, or the top 20 companies that employ the most customers. Imagine if this example comes from a large volume of records. The consistency issue then becomes exponentially more problematic.

Here is another example of data collected for a "title":



VP Sales
Vice President of Sales
Vice President Sales
Sales VP
V.P. Sales

Again, the results are inconsistent. If a company wants to develop a marketing campaign targeting sales managers, this data subset is not useful.

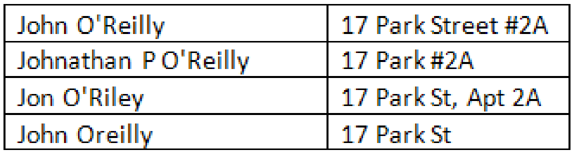
The solution is to integrate logic into the application, so that the program knows to extract data considering these "title" variations. Otherwise, marketing is going to miss potential opportunities.

Drop down boxes, combination boxes and other "choice" requirements work for data that includes a small subset of options. With most data, however, the number of possibilities is far too great and filters are rarely consistent across applications. Integrating a data standardization strategy will drastically reduce, if not eliminate, these issues, Again, this will improve the quality of data to make it more useful and actionable for data stewards, as well as generate more effective business results.

3. Matching Multiple Data Records Reduces Duplication

Many businesses face the problem of duplicate records, meaning a customer or data entity exists in a database multiple times, usually with different variations.

For example:



John O'Reilly	17 Park Street #2A
Johnathan P O'Reilly	17 Park #2A
Jon O'Riley	17 Park St, Apt 2A
John Oreilly	17 Park St

A human sees this and realizes these entries are probably referring to the same person since the names and addresses are similar. However, a computer needs the help of heuristics or other matching logic to make the same conclusion.

If an individual exists within a customer database multiple times, it creates multiple and disjointed records, as well as redundant or incomplete information. This causes, more often than not, a plethora of problems.

Possible consequences include the following:

- Customer counts are wrong. Imagine the consequences if this was a case in a business merger or acquisition when the purchaser uses a data cleanse and discovers they received 70% fewer customers than expected.
- Customer data is fragmented. This means it is often difficult to find or identify a customer and reports lack total insight. For example, fragmented data may overlook complete customer transactions, as well as different people within the same company may reach out to a customer.

As another example, if an individual exists within a customer database multiple times, it creates multiple and disjointed records, redundant or incomplete information. This may cause a poor customer experience as different people within the same company reach out to a customer or significant issues if the data is driving mission-critical decisions. Business rules to consolidate and match records can be more complex than rules that govern the matches themselves. However, once data quality tools are implemented, they reduce – if not eliminate – duplicate records.

4. Validating Third-Party Data Improves Accuracy and Enriches Data

Data can be standardized, complete and exist in a desired format, but that does not mean the entity represented by the data actually maps to a real-time or existing record.

Let us use a mailing address as an example. The standardization rule for “Street” might be “ST” and “avenue” should always be “AVE.” This does not indicate the address actually exists, or that a package will ship to that specific address. Consider, “790 Main Street, Akron, Ohio 44301.” Realistically, however, Main Street addresses end at 599.

Data quality tools from trusted, reliable sources validate and verify the address. If an address is invalid, there are different correction options. In the case of an interactive software application, the original provider of data may have an opportunity to provide a correct address.

Unless a business embeds a data verification/validation tool into an application, the deliverability of the address is uncertain.

By validating addresses, accurate deliverability is certain to happen. To enhance the information, data tools identify and correct missing or invalid elements such as a Zip Code or a Zip+4. Data verification/validation tools provide additional potentially useful data elements like latitude and longitude coordinates, county information and other postal-related information that help deliver increased value. In short, data validation tools prevent bad shipments, save wasted marketing material costs, and improve customer service/satisfaction.

Third-party data quality providers also enrich and append other types of data including phone numbers, email addresses, international addresses, business information, or any other type of information for which a master reference database exists.

5. Exploring Other Types of Data to Help Drive Results

In addition to the tools discussed thus far, there are data quality processes for numerical and non-textual information for further business value.

For instance, an elementary school wants access to birth dates in a certain geographic area for the past 12 years. Alternatively, a business wants to flag returns exceeding $100 for review. Rule-based constraint logic for transaction amounts, salaries, inventory numbers, and other business data can become part of a DQM strategy.

In addition, statistical outlier identification is of value and it is oftentimes more sophisticated. Say a business needs to pull information that meets certain criteria when compared to a whole data set, such as "99% of X."

Improving Data Quality and Measuring Success

How does a business implement these methods to make tangible and lasting improvements with their data assets?

Many solutions claim to offer data profiling, duplicate data removal, and data cleansing. To help determine the best direction to pursue, businesses need to identify low-hanging fruit. Begin by focusing on the data that is of the highest value to your business or causing the greatest concern.

Data profiling is usually the first step. Then, approach finding the right supplier. Like any other business purchase or hire, you should conduct research, speak with references, review customers, and evaluate results.

In order to prevent degradation of data over time, make the most of your investment and avoid returning to the same situation in which you started. Carefully assess the available solutions and services by weighing the features, benefits and potential ROI.

Consider utilizing cloud-based APIs to ensure real-time data quality. These solutions and services are more effective, and in many cases, less expensive. The latter is possible because APIs validate the data before it is part of a CRM system to avoid downstream data quality issues, which originally created the problem.

For DQM success, it is critical you clearly communicate your data quality strategy; educate employees about the importance of accurate, valid, real-time customer data; and share how data affects the business across the entire organization.

Industry experts also suggest implementing incentives to ensure captured data is complete and accurate. One such incentive might be to withhold commission until the salesperson accountable completes all data fields within a customer record. Another possibility could be offering a bonus to the call center representative with the highest level of data accuracy.

It is always a best practice to evaluate the value that a new initiative delivers to a business. A multi-dimensional, comprehensive data quality program is no exception. Therefore, it is equally important in DQM to determine quantifiable success metrics to measure improvement, progress and results.

Examples could be tracking, monitoring and reporting the:

- Percentage of 100% complete records
- Frequency of duplicate data
- Number of valid addresses each month

Conclusion

This white paper discussed several basic DQM approaches, methods and tools. The tools provide a great foundation for improving the quality of your organization’s core data assets, making every process, operation, decision, and communication considerably more effective.

Remember data quality is not a one-time exercise, but rather a constant, dynamic initiative that must go beyond simply the application of technology for the best possible results.

Proper application of data quality techniques contribute to business success. The journey is not a small or an easy one, but like most things, it requires taking a first step and following through over time. While the destination may change, each step forward brings more success to everyone who relies on company data.

About Informatica

Informatica Corporation (Nasdaq:INFA) is the world's number one independent provider of data integration software. Organizations around the world rely on Informatica to realize their information potential and drive top business imperatives. Informatica Vibe, the industry's first and only embeddable virtual data machine (VDM), powers the unique "Map Once. Deploy Anywhere." capabilities of the Informatica Platform. Worldwide, over 5,000 enterprises depend on Informatica to fully leverage their information assets from devices to mobile to social to big data residing on-premise, in the Cloud and across social networks. For more information, call +1 650-385-5000 (1-800-653-3871 in the U.S.), or visit www.informatica.com.

Put potential to work.™

Worldwide Headquarters, 100 Cardinal Way, Redwood City, CA 94063, USA Phone: 650.385.5000 Fax: 650.385.5500
Toll-free in the US: 1.800.653.3871 informatica.com linkedin.com/company/informatica twitter.com/InformaticaCorp